DOCUMENT RESUME

ED 388 669 TM 023 060

AUTHOR Stage, Christina
TITLE Gender Differences on the SweSAT: A Review of Studies since 1975. Educational Measurement, No. 7.
INSTITUTION Umea Univ. (Sweden).
REPORT NO ISRN-UM-PED-EM-7-SE; ISSN-1103-2685
PUB DATE 93
NOTE 33p.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Entrance Examinations; College Students; Foreign Countries; Grades (Scholastic); Higher Education; High Schools; *High School Students; *Prediction; Selection; *Sex Differences; *Test Bias; Test Content; Test Results; Test Use
IDENTIFIERS Sweden; *Swedish Scholastic Aptitude Test

ABSTRACT

The Swedish Scholastic Aptitude Test (SweSAT) has been in use as a selection instrument for higher education since spring 1977. One of the greatest problems with the SweSAT is the gender difference in results. A number of studies have been performed in order to clarify where and why these differences are found. This paper summarizes these studies and their results. The studies consist of literature studies, studies of test bias and item bias models, studies of the relations between item content and gender differences, studies of whether it is possible to predict gender differences by judgmental analyses of items, studies of gender differences in different sub-groups of test-takers and studies of the relationship between marks and test results. The 43 studies on which the review is based are listed by the following categories: literature reviews, test bias models, item bias models, relations between contents and gender differences in test results, judgments of items with regard to gender differences, gender differences in different sub-groups of test-takers, and average school marks and test results. Contains 9 tables, 1 figure, and 25 general references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 388 669

GENDER DIFFERENCES ON THE SweSAT
A Review of Studies since 1975

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
CHRISTINA STAGE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Christina Stage

BEST COPY AVAILABLE

GENDER DIFFERENCES ON THE SweSAT

A Review of Studies since 1975

Christina Stage

EM No 7, 1993

Division of Educational Measurement
Department of Education
University of Umeå
Sweden

Educational Measurement, No 7, 1993. Umeå University

Gender Differences on the SweSAT. A Review of Studies since 1975.

Christina Stage
Division of Educational Measurement, Umeå University, S–901 87 Umeå, Sweden

ABSTRACT

The Swedish Scholastic Aptitude Test, SweSAT, has been in use as a selection instrument for higher education since Spring 1977. One of the greatest problems with the SweSAT is the gender differences in results. A number of studies have been performed in order to clarify where and why these differences are found. The purpose of this paper is to give a summary of these studies and the results achieved from them. The studies consist of literature studies, studies of test bias and item bias models, studies of relations between item content and gender differences, studies of whether it is possible to predict gender differences by judgemental analyses of items, studies of gender differences in different sub–groups of testtakers and studies of the relationship between marks and test results.

The Swedish Scholastic Aptitude Test, Swe SAT, has been in use as a selection instrument for higher education since Spring 1977. During the first years only those who were eligible for higher education by being at least 25 years old and having worked for at least four years could use the SweSAT. In 1991 new rules of admittance to universities and colleges came into use and according to these new rules SweSAT is an alternative to average marks from upper secondary school and can be used by all applicants for higher education.

Since SweSAT is used as a selection test, it should rank the applicants as fairly as possible with regard to expected success in higher education. The SweSAT consists of six different subtests, all comprising multiple-choice items. The composition of the test, the number of items in each subtest and the reliability coefficients of each subtest is shown in table 1.

Table 1. The SweSAT subtests and reliability coefficients (KR_{20}).

Subtest (abbreviation)	**Number of items**	**Reliability**
Vocabulary (WORD)	30	.84
Quantitative reasoning (DS)	20	.74
Reading comprehension (READ)	24	.75
Interpretation of diagrams, tables and maps (DTM)	20	.78
General information (GI)	30	.71
English reading comprehension (ERC)	24	.84
Total	148	.95

One of the greatest problems with SweSAT – a problem which has been noted ever since before the test came into use – is the gender differences in results. A number of studies have been performed in order to clarify the differences and to clarify where and why they come into existence. The purpose of this paper is to give a summary of these studies and the results achieved from them.

The problem of gender differences is not unique for SweSAT, on the contrary gender differences in educational outcomes have been subject to extensive research during several decades. But even though educators and researchers have long been aware that differences exist between males and females regarding educational outcomes it is only recently, and very much through the widened use of SweSAT, that the problem has attracted public attention in Sweden. In the general opinion the test is unfair to females as they generally score lower than males. In Sweden where the equality between the sexes is comparatively far advanced there is no understanding for gender differences in test results, and several voices have been raised for abandoning the tests, just because

of the gender differences in results, even though public opinion in all other aspects is very much in favour of the test.

Literature reviews

The most natural starting point, when trying to understand the gender differences in results on the SweSAT, was of course the existing literature on the subject of gender differences. The first literature review within the framwork of SweSAT was finished in 1975 (Stage). In this first review the classical works were summarized, like for example Anastasi (1958), Terman & Tyler (1954) and Maccoby (1967). Since then the literature has been followed and summarized now and then: Stage (1982) covered for instance Maccoby and Jacklin's book (1974) and in a dissertation (Stage, 1985) the then recent literature was covered. A proper summary of the present situation may be borrowed from Wilder & Powell (1989):

> *"Many different tests given over a wide range of ages and educational levels reveal male-female score differences.*
> *In general the largest differences appear in tests on mathematical or quantitative ability where men tend to do better than women...*
> *The historical small advantage enjoyed by females in the verbal domain appears to have been eliminated or in some cases, reversed. (p v & p 14)*

Test bias models

In the seventies the focus was on fairness in testing and several models for fair selection were presented. These models - from Cleary's (1968) regression model through Thorndike's (1971) constant ratio model, Darlington's (1971) subjective regression model, Linn's (1973) equal probability/equal risk models, Cole's (1973) constant probabiblity model up to Petersen's (1975) expected utility model - have been described and some of them have been tried empirically on results from the SweSAT (Stage, 1978, 1985). The criterion used was average marks and the results generally were that females should be admitted to higher education on a lower test score than males. The models disagreed, however, regarding where to set the cutting score. As the criterion used was neither valid nor relevant enough these results were only regarded as illustrations. However, all the models depend on the quality of the criterion measure and it is almost impossible to measure the criterion in a way that is valid, relevant and reliable enough. Therfeore the main conclusion of the studies was that none of the strictly statistical test bias models was to be recommended for practical use and could not solve the problems of SweSAT. Perhaps the most important result of these studies of test bias models was the recognition that what constitutes a fair use of tests is not simply a technical question but is an issue which involves value judgements on fairness. And even though the idea is tempting, you can not find a strictly technical resolution to a problem which involves value judgements.

Item bias models

Another approach to the examination of test fairness was found under the heading of item bias or content bias. Item bias is at hand when single items in a test favour or disfavour some particular subgroup of the population for which the test is intended. Item bias methods detect items that are deviant i e which measure something else than the test as a whole. The main advantage with these methods is that they do not depend on an external criterion measure to the same degree since the test bias methods, as the total test score is used as criterion. Cole (1981) distinguished between two major categories of item bias models: those based on item-group-interaction and those using item-response-theory-approaches. The different models which have been presented as item bias models or item bias methods have been followed and some of them have been used to a great extent on results from SweSAT with men and women as groups. In the first study (Stage, 1980) Angoff's delta-plot-method (1972) and Scheuneman's chi-square-method (1977) were applied on two vocubalary and two general information tests. According to the first method an item is biased if it is particularly difficult for a subgroup of the population in comparison with the other items of the test. According to the second method, an item is unbiased if the probability for a correct response is the same for individuals with the same ability regardless of group membership. Of the 120 items analysed 22 were found biased by both methods; these items were further analysed with regard to specific item content in order to find out what made them biased. Five WORD-items were biased against men and of these, four were adjectives and one was a noun and three were connected to nursing. Eight WORD-items were biased against women; all were nouns and no particular subject was represented. Of the GI-items two were biased against men and they both were about children, one about a children's book and the other was about a child desease; seven items were biased agianst women and of these, three were about geography, two about wars, one politics and one labour market. The delta-plot-method has later been applied on all the subtests in SweSAT (Stage, 1990), but biased items are most often found on the WORD and GI subtests, which might seem surprising since the subtest DS and DTM are those which give rise to the greatest differences between men and women; on second thought the results are quite logical since the total test result is used as the criterion and as females generally score lower, the items on these subtest should be more difficult for females.

In some later studies (Stage, 1990, Wester, 1992) the Mantel-Haenszel item-bias-method has been applied to some subtests of SweSAT. By this method almost all the items in WORD and GI are biased which might depend on the very large groups of examinees in combination with the great diversity of the items on these subtests. There has been some problems in interpreting the results and the statement of Scheuneman (1982) is appropriate:

> *"Over the past decade a number of statistical procedures for the detection of item bias have been introduced and their validity investigated.... Relatively little work, however, has been done concerning how the statistical results are to be used once they have been obtained ... When many of us began working on the topic of bias, the isolation of items that seemed to be biased for or against a particular group appeared to be the bulk of the task. We*

> *naively assumed, as many investigators are assuming today, that a review of such items would readily reveal the source of the apparent bias, that the problem could then be easily corrected with suitable modifications or by dropping the item from the test or item pool, and that a "debiased" instrument would result." (p 180 in Berk, Ed., 1982)*

The main conclusion from the studies on item bias has been that the most fruitful approach to the understanding of gender differences in test results is to analyse all items, which differ between males and females, regarding content, format, solution strategy etc.

Relations between item contents and gender differences in test results

In 1985 (Stage, 1985) the first elaborated classification of items was made based on the results of men and women. 450 WORD–items and 450 GI–items were classified according to both item content and gender differences in result. First the items were classified according to specific item content into subject areas (by four independent persons). Then the items were categorized according to the size of the differences between men and women in response. Seven categories were used for differences in item difficulty:

1. Extremely female items: $p_F - p_M > .20$
2. Clearly female items: $.20 \geq p_F - p_M > .10$
3. Female items: $.10 \geq p_F - p_M > .02$
4. Neutral items: $-.02 \leq p_M - p_F \leq +.02$
5. Male items: $.02 < p_M - p_F \leq .10$
6. Clearly male items: $.10 < p_M - p_F \leq .20$
7. Extremely male items: $p_M - p_F > .20$

Items which were "extremely male" – 14 WORD–items and 52 GI–items – usually were about "sports, physics, geography, politics or economics". Items which were "extremely female" – 6 WORD–items and 4 GI–items – were about "home economics or diseases". Neutral items – 121 WORD–items and 67 GI–items were mainly about "biology, education, religion or literature".

The overall conclusion was that there are content areas which may be labelled as distinctly "male" or "female" but male areas are much more frequent than female areas.

The categorization of items used in this study has later been applied to all subtests of SweSAT (Stage, 1986, 1988, 1992). An example of how items of the different subtests are distributed on the seven different categories is given in table 2.

The six subtests can be ordered into three groups with regard to gender differences: group one contains the two subtests WORD and GI. These two subtests function very much in the same way: they both have items distributed on all seven categories regarding p–differences. The two subtests are also similar as to the type of test they

represent; they both demand previous knowledge (about the meaning of a word or about some speical fact) if the correct answer is to be produced. It is also evident that the subject content of the items is the main reason for the resulting p-differences between males and females on these subtests.

Table 2. Distribution (in per cent) of items on the seven categories described above.

Subtest	Female			Neutral	Male			Number
	1	2	3	4	5	6	7	of items
WORD	2	13	33	23	21	6	2	180
DS	0	0	1	5	43	50	1	120
READ	0	1	11	43	41	3	1	144
DTM	0	0	0	7	50	42	1	120
GI	1	5	17	19	31	21	6	180
ERC[1]	0	0	0	13	75	10	2	48

The second group of subtests consists of DS, DTM. These are the two subtests which may be said to be the most quantitative. On these tests practically all the resulting p-differences are in favour of males regardless of subject content and also regardless of mathematical content (see below).

It is remarkable that the new subtest ERC seems to belong to the same group as DS and DTM. ERC is a verbal subtest and could be expected to be more similar to the subtest READ. As ERC has only been used twice the knowledge about it is very limited so far.

The third group consists only of the subtest READ. In this subtest the differences are spread from the category "clearly female" to "extremely male". A closer study of the specific content of this test supports the conclusion drawn from the studies of GI-items that some subject areas are more favourable for males and some are more favourable for females, but there are more areas which favour males. (see Stage, 1986 and 1988).

On the subtest DS some special studies have been performed. In one study (Stage, 1987) 160 DS-items were analyzed with regard to subject content as well as the kind of mathematical solution demanded. In tabel 3 the results for different contents is presented.

---

1 This subtest has only been in use since Spring 1992.

Table 3. Different content areas, average differences in p-values between males and females, range of the differences and number of items with a difference less than .06 and number of items with a difference larger than .15.

Content	Number of items	Average $p_M - p_F$	Range of diff	Number .00 -.05	Number > .15
People	32	.11	.00-.19	4	8
Taxes/salaries/prices	27	.12	.04-.21	4	7
Communications	16	.10	.03-.22	5	2
Abstract content	13	.10	.01-.18	2	2
Animals	12	.09	.01-.13	2	-
Energy/metals	12	.12	.04-.27	2	2
Home economics	12	.09	.03-.18	3	1
Environment	6	.11	.03-.16	1	1
Time	4	.09	.04-.23	1	1
Sports	3	.14	.09-.17	-	1
Others	23	.09	.03-.23	6	1
Total	160	.10	.00-.27	30	25

The content areas that gave rise to the largest differences were sports, taxes/salaries/-prices and energy/metals, but no area was found to cause small or no differences. The same conclusion was drawn from the study of mathematical operations; problems containing indices seemed to cause the largest differences but no operation was found that did not give rise to differences.

A somewhat different study of the importance of subject content in the items of DS was conducted by Henriksson, Stage & Lexelius (1986). Items that were known to give large differences in favour of men were changed with regard to subject content but not with regard to mathematical solution i.e. an item where the problem was to decide "the share of the USA in the total capacity of the nuclear reactors of the world" was changed to "the share of Stockholm in the Swedish capacity of child-daycare for children". The only safe conclusion from the study was that the items became easier (for both males and females) when the content was changed in the female direction, but the differences between men and women did not change.

On the subtest DTM one study was conducted (Wester-Wedman, 1992) with the aim to compare the results of males and females on an ordinary multiple-choice-test and a parallell test with open-ended questions. The hypothesis was that the differences would be smaller on the test with open-ended items since a lot of studies have shown that the multiple-choice format favours males (see f ex Murphy, 1982). The conclusion was, however, that the changed item-format did not cause any decrease in the differences between males and females.

Judgements of items with regard to gender differences

One problem when composing a test is that there are no clear guidelines for what content is male and what is female. In the study on GI-items described above the results on every single item in nine tests were classified with regard to differences between men and women and subject content.

The average results of men and women on the subtest of GI differ between one and two points. Hence the difference is a small but very consistent one and the general opinion does not really accept that there should be a "true" difference between men and women regarding general information.

The number of items from different subject areas of the GI test is determined in proportion to the size of the different faculties at Swedish universities, which means that 8 – 10 items are from the technical or natural sciences sector, 10 – 12 items are about social sciences, 6 – 7 from the humanities sector, 1 – 2 items are from the sector of education and training and 3 – 4 items are from the sector of nursing and medical care.

When 450 GI-items were classified according to gender differences and subject the results were as shown in table 4.

Table 4. Subject areas, (number of items from each area), results regarding gender differences and average differences in p-values between men and women.

Subject area (number of items)	Category							$p_M - p_F$
	Female		Neutral			Male		
	1	2	3	4	5	6	7[2]	
Technical/natural sciences (99)	2	9	16	16	24	19	13	.06
Social sciences (169)	–	3	8	17	55	55	31	.11
Education or training (35)	–	3	5	2	17	8	–	.04
Humanities (106)	–	6	14	24	34	21	7	.05
Nursing and medical care (41)	2	5	17	8	6	2	1	–.03
Total (450)	4	26	60	67	136	105	52	.07

[2] 1 and 7 are extremely female/male; 2 and 6 are clearly female/male; 3 and 5 are slightly female/male and 4 is neutral.

It was evident from the results in table 4 that the division of items in accordance with the university sectors was too broad to give any useful information about which content is easier for men and which is easier for women. All sector areas, but Nursing and medical care seem to favour men to a slight degree, even though the differences are small. In all sectors, however, the items are distributed on both the male and female side, though generally there are more items on the male side. A further division of the items according to subject content was possible, however, and the results from this division are shown in table 5.

Table 5. Subject areas, (number of items from each area), results regarding gender differences and average differences in p–values between men and women.

Subject area (number of items)	**Category**							$p_M - p_F$
	Female			**Neutral**		**Male**		
	1	**2**	**3**	**4**	**5**	**6**	**7**	
Technical/natural sciences:								
Physics (26)	–	–	–	1	8	8	9	.19
Chemistry (21)	–	–	1	4	5	7	4	.12
Biology (31)	1	4	8	7	8	3	–	–.01
Home economics (21)	1	5	7	4	3	1	–	–.06
Social sciences:								
Geography/history (29)	–	–	–	5	9	7	8	.12
Laws/economics (50)	–	2	4	7	13	19	5	.09
Swedish politics (36)	–	–	1	4	17	10	4	.10
International politics (54)	–	1	3	1	16	19	14	.14
Education or training:								
Labour market (21)	–	1	1	1	11	7	–	.06
Education/administration (14)	–	2	4	1	6	1	–	.00
Humanities:								
Religion (9)	–	1	2	2	3	1	–	–.01
Literature (55)	–	2	7	12	20	12	2	.06
Art/music/film (33)	–	3	5	9	10	5	1	.02
Sports (9) –	–	–	–	1	1	3	4	.16
Nursing and medical care:								
Diseases (20)	2	4	8	3	3	–	–	–.07
Methods of treatment (13)	–	1	5	4	1	1	1	.00
Health administration (8)	–	–	4	1	2	1	–	.00
Total (450)	4	26	60	67	136	105	52	

When the five sectors were taken as a whole no great differences were found between the mean differences in p-values between males and females. The area of social sciences was the largest with a mean difference of .11. When the sectors were divided into subareas, however, the pattern changed. Within the technical and natural sciences area there is quite a difference between physics and home economics for example. Also the rank order between difference sizes changes. There is still, however, quite some dispersion of the item results within subjects. No subject is homogeneous with regard to differences.

One method which has often been recommended for determining the content validity or to assure that a test is free from bias is to let subject experts study the items and determine their relevance. Tittle (1975) recommended a similar method to ensure that items are fair to different groups:

> *"Item pools can be judged by having subgroup members predict performance. Judges would rate each item as to which member of a set of paired subgroups would get the item right more often or if the rater expected there would be no difference in performance." (p. 91)*

Berk (1982) gave a similar recommendation in a step-by-step procedure for trouble-shooting test bias, in which step three, four and five are:

3. *Select a panel of reviewers according to the sex and ethnic composition of the target population.*
4. *Devise procedures and instruments (e.g. checklists) for judgemental analysis.*
5. *Conduct judgemental analysis. (p. 6)*

The judgemental review in the procedure suggested by Berk was, among other things, intended to assure fair representation of the experiences of different groups.

After the results are obtained it very often seems quite obvious at least regarding verbal items or items on general information, which items are easier for males and which are easier for females. Unfortunately this is not always that obvious before the results are available.

In a study performed by Wedman & Stage (1983) on students from upper secondary school, the task was to rate 35 items on general information regarding expected gender differences in results. The items had been chosen from a bank of 150 items for which results of men and women were available. All existing female items were chosen but that resulted in only seven in all and to these were added 17 male and 11 neutral items. 153 boys and 162 girls took part in the study. The results were that for the male items 31 % of the ratings were correct, 62% were ratings for the items as neutral and 7% as female; of the female items 50% of the ratings were correct, 49% as neutral and 1% as male. 72% of the neutral items were rated as neutral. There was no difference between boys and girls regarding the correctness of ratings. The conclusion of the study

was that it seems to be easier to identify female than male items, which more often were rated to be neutral.

One objection which was raised against this study (Emanuelsson, 1985) was that the larger number of male items than female items coaxed the subjects to level the numbers and therefore rate male items as neutral. Another objection was that students from upper secondary school can not be expected to be able to do this kind of ratings.

The aim of another study (Stage, 1987) was to examine to what extent a group of experts regarding test construction and test evaluation is able to determine which items in general information are easier for men and which are easier for women.

In this study a test on GI was compiled of 10 male, 10 female and 10 neutral items. Items were categorized as male and female if the differences in p-values had been .09 or .10. No items with larger differences were used as the wish was to avoid very extreme items. Neutral items were items without any differences in p-values.

The mean of the 30 items was 18.09 for earlier examinees regardless of gender. With these restrictions on p-value-differences between men and women it was not possible to compile a test with the usual distribution on different subject content (see above). The test consisted of 7 items from the technical or natural sciences area, 2 from the social sciences area, 17 from the humanities area and 4 from the area of nursing and medical care. Hence the number of items from the area of humanities was larger and the number of items from the area of social sciences was smaller than in an ordinary subtest on general information. The items were compiled in random order to a subtest on general information.

19 persons, twelve males and seven females, all of which had long experience from either test construction or test evaluation, were asked to rate the 30 items regarding expected gender differences in result.

The subjects were presented the subtest and were asked to rate each item in one of the three categories - male, female or neutral - regarding outcome for a group of applicants for higher education. They were not informed of the number ot items from each category, but were told that the mean results were the same for men and women.

The result was that 14.4 of the ratings were correct, i.e. a bit less than half of the items were rated correctly. The best individual result was 18 correct ratings and the worst individual result was 10 correct ratings. Table 6 shows the ratings for the three groups of items.

Table 6. Ratings of items regarding expected gender differences in results.

Category	Mean of Ratings			
	Female	Neutral	Male	Total
Female	**5.94**	3.32	0.74	10.0
Neutral	3.52	**3.80**	2.68	10.0
Male	0.89	4.37	**4.74**	10.0
Total	10.35	11.49	8.16	30.0

There is a significant difference between the three categories. In this study the neutral items were most difficult to rate, but still the female items were easiest to rate. 59% of the female items were rated correctly; 38% of the neutral items and 47% of the male items were rated correctly. The conclusion remains, however, that it is difficult to rate items correctly regarding gender differences in result, but it is easier to rate female items than male or neutral items.

Gender differences in different sub-groups of testtakers.

In 1991 the rules for admittance to tertiary education were changed and already in 1990 the group of examinees had changed. Generally this new group was younger and had higher education than the examinees at earlier test administrations. Since 1990 it has been possible to match the groups of men and women in a more meaningful way regarding age and education and several studies have aimed at determining the importance of age and education for the gender differences in test results. (Bränberg et al, 1990; Stage, 1991, 1992a,b,c).

The conclusion from the first study (Bränberg et al, 1990) was:

> *"The results indicate rather genuine differences in every variable studied. Test takers with a higher education obtain higher mean score than those with a lower education and older testtakers obtain higher mean score on the subtests WORD and GI than younger persons. The mean test score for men is higher than the corresponding score for women, even if differences in education and age are controlled for." (p 189)*

And the conclusion from another study (Stage, 1992) was very similar:

"The results, however, showed that even though age as well as education had influence on the test results, no real difference was found between younger and older examinees regarding gender differences in the test results."(p 223)

It has been noted, however, that even though matching of males and females regarding length of education has a limited effect on the differences in results, matching regarding content of education has a greater effect.

The largest group of examinees in 1992 was applicants who had finished three-year upper secondary school[3]. In Autumn 1992 the mean difference between males and females with finished three-year upper secondary school was 8.8 points. There are, however, five different course programmes in upper secondary school and the difference between the course programme that score highest and the one that score lowest is 13.6 points. In table 7 the number of males and females on the five different course programmes are given as well as their results on each subtest.

Table 7. Results Autumn 1992 for males and females, on the six subtests and the total test, distributed on different course programmes in upper secondary school.

Course	H		S		Ec		N		En	
Sex	M	F	M	F	M	F	M	F	M	F
Number	*281*	*1594*	*1965*	*3176*	*3193*	*3190*	*3088*	*2356*	*4432*	*1101*
Subtest										
WORD	23.5	22.7	21.2	20.3	19.8	19.2	22.0	21.5	20.3	20.2
DS	11.8	10.8	12.8	11.7	13.0	11.6	15.1	13.9	14.6	13.9
READ	16.9	15.2	16.3	15.1	15.6	14.3	17.6	16.7	16.6	15.8
DTM	14.9	13.3	15.7	14.1	16.2	14.2	17.5	16.3	17.3	16.6
GI	20.5	19.0	19.4	17.9	18.4	16.8	20.3	19.6	18.5	17.7
ENG	20.4	19.0	19.4	17.9	18.4	16.8	20.3	19.6	18.5	17.7
Total	108.0	99.7	105.5	97.6	101.8	93.5	114.0	108.1	107.6	103.2

As may be seen in table 7 the content of education seems to have great influence on the results. Student who have attended the natural science study course score higher than all the other students and that is true for males as well as females, even though the females score lower than the males from the same course. It may also be noted that the gender

3 After nine years compulsory, integrated education the students can choose between different course programmes in upper secondary school. Five of the course programmes prepare for higher education and they all last for three years; these different course programmes are: the Humanities course (H), the Social Science course (S), the Economic course (Ec), the Natural science course (N) and the Engineering course (En), which has an optional fourth year.

differences are smaller on the natural science and the engineering courses than they are on the other study courses.

The overall conclusion that men have greater general knowledge gives the impression that men on the average have somewhat higher scores than women on all items dealing with general information. The analysis of the results on individual items reported on page 7 & 8 demonstrated that this is not the case. On 90 items out of the 450 GI–items studied women had on the average higher scores (statistically significant) and on 67 items no differences between men and women could be observed. The conclusion should rather be that within certain areas of general information men have greater knowledge, within other areas women have greater knowledge and within some areas no differences exist.

The latter conclusion is supported by the results from another study (Stage, 1985). There a comparison was made between men and women obtaining equal scores on the GI–subtest: a group of males whith a total score of 18 was compared to a group of females with the same total score. When the results on individual items were compared for these groups of males and females it was found that even between these equalized groups there were substantial gender differences in results on separate items. Items dealing with health care and home economics were easier for females whereas items dealing with sports and economy were easier for males. Only 14 of the 60 items studied were of the same difficulty for males and females, even though they in the normal way of interpreting a test score ought to have equal general information.

The conclusion that men and women have equal knowledge of vocabulary is not correct either. Only for 121 out of the 450 WORD–items described on page 4 were the average proportions of correct answers equal for males and females. On 112 items females had on the average higher scores and on 212 items men had higher average scores. The similarity of the total results is due more to the balancing of items than to males and females having the same vocabulary knowledge.

The results from this study demonstrate still further that males and females obtain their scores in different ways. In the same study as described above a comparison on item level was made between those males and females who had a total score of 18 on the two WORD–subtests. Only 11 of the 60 items had the same difficulty for males and females, whereas nine items belonged to category two (see page 4) and 13 items belonged to category six even for these groups of males and females who ought to have the same vocabulary. A clear tendency with respect to words could also be observed in that among the female words were: placenta, assiduous, litany, chimera, condole, perennial and intrepid and among the male words were: synchronous, patronize, undermine, polemics, hypothesis, decadence, evict, grumble, sonorous and ratify. Thus, these words differentiate very much between groups of males and females even when the two groups are considered to have equal vocabulary knowledge.

Average School Marks and Test Results

As was mentioned earlier the largest group of test-takers since 1990 has been those who have finished three-year upper secondary school: they constitute about 65% of the total group of test-takers. As has also been mentioned earlier there are five different study course programmes in three-year upper secondary school and these courses differ in test results in a very systematic way. As there is in principle two ways to be selected for higher education: Average marks from upper secondary school or results on SweSAT, it has been of interest to compare the two selection instruments especially for males and females. (Stage, 1992a, b, c).

Marks from the five different course programmes in upper secondary school are regarded as equivalent which means that when applying for higher education only the average mark of a student is important and not which course programmes he/she has attended. There is a considerable interaction between gender and choice of academic course programme as well as between course programme and test results. In table 8 the entrance and leaving marks are shown for students who finished upper secondary school in 1991 and also the marks for those who took part in the SweSAT in 1991.

Table 8. Average entrance and leaving marks for students who finished three-year upper secondary school 1991, and leaving marks for those who took part in the SweSAT in 1991, distributed by sex and course programme. *N* is the number in each group.

	Entrance marks			Leaving marks			Testtakers' marks		
	M	**F**	**Total**	**M**	**F**	**Total**	**M**	**F**	**Total**
Course									
H	3.62	3,72	3.70	3.28	3.40	3.39	3.52	3.59	3.58
N	*426*	*3052*	*3478*	*371*	*2864*	*3235*	*95*	*891*	*986*
S	3.73	3.83	3.80	3.24	3.48	3.41	3.45	3.60	3.56
N	*2212*	*5511*	*7723*	*2396*	*5531*	*7927*	*942*	*2651*	*3595*
Ec	3.54	3.71	3.65	3.18	3.30	3.25	3.46	3.54	3.51
N	*4942*	*8018*	*12960*	*5047*	*7503*	*12550*	*1496*	*2263*	*3759*
N	3.95	4.19	4.07	3.70	3.80	3.75	3.78	3.82	3.80
N	*3699*	*3915*	*7614*	*3309*	*3538*	*6847*	*2350*	*2667*	*5017*
En	3.72	4.16	3.82	3.29	3.40	3.31	3.54	3.52	3.53
N	*9708*	*2714*	*12422*	*8524*	*2168*	*10692*	*3379*	*1126*	*4505*
Total	3.72	3.87	3.80	3.32	3.45	3.39	3.58	3.64	3.61
N	*20987*	*23210*	*44197*	*19647*	*21604*	*41251*	*8262*	*9598*	*17860*

The pattern of educational choice is very different for males and females. More females have chosen the humanities, the social science and the economic course programmes while more males have chosen the engineering course programme. On the natural science course, however, there are as many females as males; this course programme was formerly dominated by males.

The average marks of females (at entrance as well as at leaving) are generally higher than those of males within all different course programmes. The sizes of these differences are, however, different on the various course programmes. With regard to entrance marks the differences in favour of females are greatest for the natural science and the engineering course programmes but on these two programmes the differences in favour of females are smallest at the end of the studies.

In table 9 the test results are shown, distributed on males and females and different course programmes, for the testtakers in 1991 who finished their upper secondary school education in 1991.

Table 9. Mean scores and number of testtakers from different course programmes in upper secondary school and the proportion of testtakers in per cent of each group.

	Average test score			Testtakers in per cent		
	M	**F**	**Total**	**M**	**F**	**Total**
H	98.1	88.8	89,7	26	31	30
N	*95*	*891*	*986*			
S	97.6	89.1	91.3	39	48	45
N	*942*	*2 651*	*3 593*			
Ec	94.5	86.0	89.4	30	30	30
N	*1 496*	*2 263*	*3 759*			
N	109.8	101.4	105.3	71	75	73
N	*2 350*	*2 667*	*5 017*			
En	99.8	93.7	98.3	40	52	42
N	*3 379*	*1 126*	*4 505*			
Total	101.4	92.3	96.5	42	44	43
N	*8 262*	*9 598*	*17 860*			

A comparison between the different course programmes shows that testtakers from the natural science course have the highest scores followed by students from the engineering

course programme; students from the economic course programme have the lowest scores. This has always been the ranking between the course programmes regarding results on the SweSAT (Stage, 1992). Note, however, that this was not the ranking regarding average leaving marks where the students from the engineering course programme were second last.

In accordance with similarities with regard to subjects studied the course programmes may be grouped as follows: H, S and Ec constitute one group and N and En constitute the other. In figure 1 the results are shown for males and females from these groups of study course programmes in different intervals of average marks.

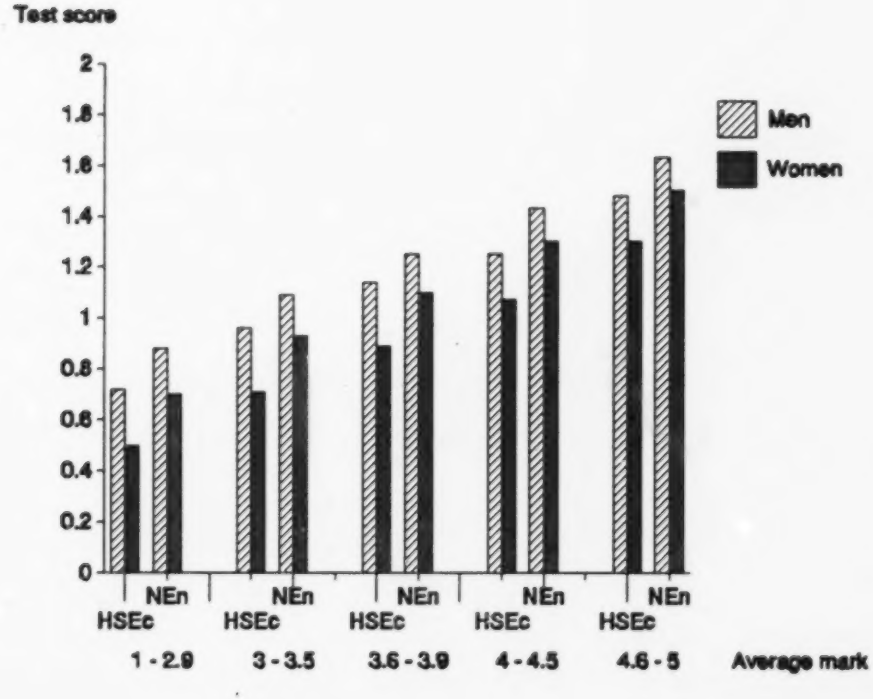


Figure 1. Standard test scores for men and women from different groups of course programmes in upper secondary school and in different intervals of average marks.

The differences between males and females are notable also within the two groups of study programmes. However, the differences between the two groups of course programmes are almost more striking. The differences betwen males and females within the groups of course programmes is very much of the same size as the differences between the groups of course programmes. It is evident that the course programme taken in upper secondary school has greater impact on the test results than the level of average marks.

Within the same course programme in upper secondary school girls get lower average test scores than boys, though girls who have taken the natural science or the engineering course programme have as high or even higher scores than boys who have taken other course programmes. The content of education is of decisive importance, but boys seem somehow to be able to make better use of their education than girls do.

It is interesting to note that the rank order of females from different course programmes regarding entrance marks to upper secondary school is exactly the same as the rank order regarding test results (N, En, S, H, Ec), while the rank order regarding leaving marks is quite different (N, S, H, En, Ec). The entrance marks are given when all students have attended the same curriculum and test results are given for the same performance, while leaving marks are given after different course programmes. All students lose in average marks between entering and leaving upper secondary school and this loss is on the average .41. There is, however, a significant interaction between course programme and gender with regard to the amount lost; females lose .76 on the engineering course programme, while males lose .43 on the same course. On the other hand males lose .49 on the social science course programme while females lose only .35. This interaction between gender, course programme and average marks makes it easy to agree with Rigol:

> *"Relatively little research has been conducted about differential grades, but just as objective test results have been scrutinized, so should grades be examined." (p vi)*

Concluding remarks

The starting point for all these studies have been the average differences between males and females which have been observed in test results. The initial aim was to examine whether it could be considered fair to use test results despite these average differences. Gradually the aim has been changed towards examining to what extent the content in test items is related to differences in results and whether the differences are consistent for different groups of men and women. All these questions can be viewed as questions regarding the validity of the tests, since the determination of the validity can be regarded as an overall evaluation of all the conclusions that can be drawn from test scores. All the studies reported can thus be seen as contributions to the validation of the tests under study.

Assuming that based on the theory of validity it is possible to establish that the optimal selection would entail accepting boys and girls at different test scores, political values would still be needed to reach conclusions regarding the appropriateness of such a procedure. Even within the framework of the theory of validity, objections could be raised against using different selection scores in this specific situation as the criterion is very dubious.

There may be several reasons for items to function differently for different groups of examinees. There may be differences between groups as regards experience and training

or as regards values or cultural background and these differences may lead to groups achieving differently for different items. Scheuneman (1982) has suggested a crosstable for items which function differently for groups. One basis of division is whether or not items are related to what one wants to measure by the test. The other basis of division is the reason for groups achieving differently, which can be either that they have different experiences/training or that they have different values/cultural background. Theoretically it is easy to accept that only those items, which are not related to what one wants to measure by the test, are biased and ought to be eliminated. Items which are related to what one wants to measure can not be labelled biased since differences in results on these items are part of what one intends to measure by the test and therefore ought to be seen as valid differences.

In theory the problem of items resulting in average differences between groups is easily solved. One ought to eliminate those items which are not clearly related to what one wants to measure and retain those items which have a clear such relation. The problem remains, though, in the practical application since it is often very difficult to determine whether an item is clearly related or not to what one wants to measure. When group differences have been demonstrated in results on individual items it is necessary to make a subjective judgment for each item whether the content of that item is related or not to what one intends to measure.

Normally, one does not accept that any differences in knowledge exist between men and women and therefore the aim is often in test construction that the tests should result in approximately the same average scores for men and women. Such a result can be obtained by empirical balancing, i.e. the items are selected in such a manner that either only those items are included for which the differences at the try-out have been small or if items are included which give large differences in favour of one group, such items are matched by items giving similar differences in favour of the other group.

All efforts to balance tests with respect to the results of groups imply that one knows in which manner the content influences the applicability of the test. Further, the balancing is based on the conviction that there are no group related differences in the underlying ability which the test is intended to measure.

It would be quite possible to construct tests on vocabulary and general information where women on the average would achieve as well as or even better than men. However, the composition of above all the tests on general information would differ from that which has been stipulated for the GI subtest of SweSAT (see p 7). It is difficult to foresee in what manner such a change in composition would affect the applicability of the tests. Given present social values it would presumably not be seen as an equally obvious indication of general information to interpret a washing instruction or to recognize a potted plant (two typically female items) as it would be to know Portugal's former colonies or to know the most common direction of Swedish rivers (two typically male items). Balancing would presumably affect the "face validity" of the tests negatively.

The main requirement on the tests is that they should in some general meaning be predictors for higher education. Test scores should have a positive connection with

success in higher studies. If the tests were adjusted in a "female" direction more females would probably achieve higher scores, but the risk is that those scores would have a poorer relation to success in studies. The tests have to be in line with existing values if they are to serve as predictors. If one tried to adjust the tests to some sort of desirable future values it is uncertain how they would function today.

One would rather require test constructors and those responsible for tests to adjust the tests to existing values. Wood (1978) describes the difficulties in the following manner:

> *"Examiners cannot be held responsible for the existence of sex differences and that these differences are poorly understood. But they can be held responsible for favouring one sex unduly through a reluctance to recognise that there is such a thing as gender differences" (p. 164).*

Tests constructors have the responsibility to make sure that no items are accepted which one-sidedly favour life-experience which is particular for one sex, but not to construct tests in a manner which go against existing values. It is a question of balancing two in a certain sense contradicting requirements.

All studies which contribute to the explanation of how a test score can/ought to be interpreted may be regarded as steps towards construct validation of the test. When construct validity is determined one has to look for information which conforms with or contradicts the expectations which are generated from the concepts one intends to measure. The relation between the test score and an external criterion can be used as one of many indications of construct validity.

One indication of lacking construct validity is if scores for different groups have different structures. It is mainly in this connection the item bias methods have been introduced. Item bias methods aim at discovering items which function deviantly for some subgroup of the examinees.

An early but still common method for studying the construct validity of a test and at the same time determine possible item bias is to judge the items subjectively. It is important though to keep apart two frequent types of evaluation.

The first type of evaluation has mainly to do with that kind of validity which is normally called face validity. A lack of face validity is described by Cole (1981) in the following manner:

> *"Facial bias would occur when particular words or item formats appear to disfavor some group whether or not they in fact have that effect. Thus an instrument using the male pronoun "he" throughout or involving only male figures in the items would be facially biased whether or not such uses affected the scores of women" (p. 1073).*

Many studies have been devoted to comparing male and female agents respectively in textbooks as well as in test items (see e.g. Lockheed, 1974 or Tittle, McCarthy & Steckler, 1974). In a number of cases it has been established that men and male

pronouns are more frequent than women and female pronouns and also that often the descriptions of male and female activities and roles are prejudiced and stereotyped. This type of bias in test-items, which is easy to detect and act upon (it is not found in the SweSAT), does not seem to influence gender differences in results (Dwyer, 1975, 1979, Donlon et al, 1979). This is commented by Dwyer (1979):

> *"While one may reasonably assume that such materials alter test-taking motivation for some individuals, and that such materials may also have a subtle and long-term adverse impact, there is no research to date indicating that sexist practices have any observable effect on item or test psychometric characteristics, or that they affect group score in any way" (p. 347).*

It is important to keep apart the type of evaluation described above from the other type of evaluation which implies that the items are studied with respect to a specific content. Those studies where the content of items has been studied with respect to the results of men and women have demonstrated certain relations.

If group differences are to be used for validation purposes one condition is that knowledge exists of what these group differences look like and when they can be expected to appear. All of it can be seen as a reciprocal process. All studies summarized here contribute, however, to the information of how test scores can be or ought to be interpreted.

Cleary (1991) has pointed out that:

> *"Interpretation of the gender difference data are subject to a number of caveats:*
>
> ** Gender differences do not imply nonoverlapping distributions. The distributions of boys and girls are more similar than different...*
> ** The data available for analysis have been subject to selection factors. Some of these we know, e.g.,.....but there may be additional factors of which we are unaware. (p 54).*

It may be concluded that in spite of all the studies performed the knowledge of differences between men and women with respect to intellectual achievements is still rather poor. The same can be said of the knowledge of similarities between men and women in the same respect. In those cases where differences in achievement have been demonstrated it has generally been a question of small differences. In those cases where no differences have been demonstrated nor have similarities in a real sense been demonstrated.

In a study directed towards differences in test achievements between men and women there is an inherent risk of laying too much stress on the differences at the expense of the similarities in achievement which also exist. In fact, the similarities in achievement between men and women are considerably greater than the differences. It is often expressed by the observation that the variance within groups is greater than the variance between groups, i.e. the difference between the best and poorest achieving men and the

corresponding difference for women are greater than the average differences between men and women.

It is also important to remember that the students who take the SweSAT are a selected group, i e it is by no means representative or random groups of males and females. There has actually been selection on several levels: when admitted from compulsory to upper secondary school, when choosing course programme in secondary school and when choosing to take the test. There also seems to have been a harder selection of males than of females.

REFERENCES

Anastasi, A. (1958). *Differential Psychology*. Boston: Allyn and Bacon.

Angoff, W. & Sharon, A. (1972). Patterns of Test and Item Difficulty for Six Foreign Language Groups on the Test of English as a Foreign Language. *Educational Testing Service: Research Bulletin, 2.*

Berk, R. (1982). (Ed.) *Handbook of Methods for Detecting Test Bias*. Baltimore: The Johns Hopkins University Press.

Cleary, T. (1968). Test Bias: Prediction of Negro and White Students in Integrated Colleges. *Journal of Educational Measurement, 5, 115-124.*

Cleary, A. (1991). Gender Differences in Aptitude and Achievement Test Scores. In Sex Equity in Educational Opportunity, Achievement and Testing. *Invitational Conference Proceedings. ETS, 51-90.*

Cole, N. (1973). Bias in Selection. *Journal of Educational Measurement, 10, 237-255.*

Cole, N. (1981). Bias in Testing. *American Psychologist, 36, 1067-1077.*

Darlington, R. (1971). Another Look at "Cultural Fairness". *Journal of Educational Measurement, 75, 71-82.*

Donlon, T., Ekstrom, R. & Lookheed, M. (1979). The Consequences of Sex Bias in the Content of Major Achievement Test Batteries. *Measurement and Evaluation in Guidance, 11, 202-216.*

Dwyer, C. (1975). Test Content and the Determination of Sex Differences in Reading. Paper presented at the Annual Meeting of AERA in Washington, April, 1975.

Dwyer, C. (1979). The Role of Tests and Their Construction in Producing Apparent ex-Related Differences. In Wittig, M. & Petersen, A. (Eds.) *Sex-Related Differences in Cognitive Functioning*. New York: Academic Press.

Linn, R. (1973). Fair Test Use in Selection. *Review of Educational Research, 43, 139-161.*

Maccoby, E. (1967). (Ed.) *The Development of Sex Differences*. London: Tavistock.

Maccoby, E. & Jacklin, C. (1974). *The Psychology of Sex Differences*. London: Oxford University Press.

Murphy, R. (1979). Sex Differences in Examination Performance: Do these Reflect Differences in Ability or Sex-Role Stereotypes? In Harnett, O., Boden, G. & Fuller, M. (Eds.) *Women. Sex-Role Stereotyping*. New York: Tavistock Publications Ltd. 159-167.

Murphy, R. (1982). Sex Differences in Objective Test Performance. *British Journal of Educational Psychology, 52, 213-219.*

Petersen, N. (1975). An Expected Utility Model for "Optimal" Selection. Iowa Testing Programs Occasional Paper, No 10.

Rigol, G. (1989). Introduction to Wilder, G. & Powell, K. Sex Differences in Test Performance: A Survey of the Literature. *College Board Report, No 89-3.*

Scheuneman, J. (1977). Latent Trait Theory and Item Bias. Paper presented at the Third International Symposium on Educational Testing. Leyden, the Netherlands, 1977.

Scheuneman, J. (1982). A Posteriori Analyses of Biased Items. In Berk, R. (Ed.) *Handbook of Methods for Detecting Test Bias*. Baltimore: The Johns Hopkins University press. 180-198.

Terman, L. & Tyler, L. (1954). Psychological Sex Differences. In Carmichael, L. (Ed.) *Manual of Child Psychology*. New York: Wiley. 1064-1114.

Thorndike, R. (1971). Concepts of Culture Fairness. *Journal of Educational Measure ment, 6*, 63-70.

Tittle, C. (1975). Fairness in Educational Acievement Testing. *Education and Urban Society, 8, 86-102.*

Wilder, G. & Powell, K. (1989). Sex Differences in Test Performance: A Survey of the Literature. *College Board Report No 89-3.*

Wood, R. (1978). Sex Differences to Answers to English Language Comprehension Items. *Education Studies, 4, 157-165.*

REFERENCES: Studies on Swe SAT

Literature reviews

Stage, C. (1975). Intellektuella könsdifferenser. *Spånor från Spint, 2*, Pedagogiska institutionen, Umeå universitet.

Stage, C. (1982). Intellektuella kösskillnader. En litteraturöversikt. I Larsson, K. (red) *Skola språk och kön*. Lund: Studentlitteratur. 7–22.

Stage, C. (1985). Könsrelaterade skillnader i kognitiv förmåga. I Stage, C. *Gruppskillnader i provresultat*. Akademisk avhandling nr 17, Pedagogiska institutionen, Umeå universitet. 7–19.

Test bias models

Stage, C. (1976). Metoder att bedöma testrättvisa. *Spånor från Spint, 7*, Pedagogiska institutionen, Umeå universitet.

Stage, C. (1978). Några aspekter på rättvisa vid användning av test. *Pedagogiska rapporter, 65*. Umeå universitet.

Stage, C. (1985). Test-bias. I Stage, C. *Gruppskillnader i provresultat*. Akademisk avhandling nr 17, Pedagogiska institutionen, Umeå universitet. 33–54.

Item bias models

Stage, C. (1979). Könsskillnader i ordkunskap. *Pedagogiska rapporter, 68*, Umeå universitet.

Stage, C. (1980). Att studera testuppgifter som fungerar olika för män och kvinnor. *Pedagogiska rapporter, 87*, Umeå universitet.

Stage, C. (1984). Könsskillnader i allmänorientering. *Pedagogiska rapporter, 1*, Umeå universitet.

Stage, C. (1985). Item-bias. I Stage, C. *Gruppskillnader i provresultat*. Akademisk avhandling nr 17, Pedagogiska institutionen, Umeå universitet. 55–82.

Stage, C. (1990). Mantel-Haenszel-analys av ORD, AO och LÄS. 90:A. Stencil, Pedagogiska institutionen, Umeå universitet.

Stage, C. (1991). Gruppdifferenser och bias. *PM från H-gruppen, 39*, Pedagogiska institutionen, Umeå universitet.

Wester-Wedman, A. (1992). Lösningsstrategi i DTK-provet. En studie av relationen lösningsstrategi och uppgiftsbias avseende kön hos uppgifter i DTK-provet. *PM från H-gruppen, 55*. Avdelningen för pedagogiska mätningar, Umeå universitet.

Wester, A. (1992). Item Bias with Respect to Gender Interpreted in the Light of Problem-solving Strategies. Paper presented at the IAEA 18th Annual Meeting in Dublin, September,1992.

Relations between contents and gender differences in test results.

Stage, C. (1976). Hur könsdifferenser i problemlösning kan påverkas av probleminnehåll. *Spånor från Spint, 6*, Pedagogiska institutionen, Umeå universitet.

Stage, C. (1982). Könsskillnader i ordkunskap. I Larsson, K. (red) *Skola språk och kön*. Lund: Studentlitteratur. 108-118.

Stage, C. (1984). Könsskillnader i resultat på 450 allmänorienteringsuppgifter. *Pedagogiska rapporter, 2*, Umeå universitet.

Stage, C. (1985). Analys av ordkunskaps- och allmänorienteringsuppgifter utifrån könsskillnader i resultat. I Stage, C. *Gruppskillnader i provresultat*. Akademisk avhandling nr 17. Pedagogiska institutionen, Umeå universitet. 83-118.

Stage, C. (1986). Könsskillnader i resultat på sex högskoleprov. *PM från H-gruppen, 9*, Pedagogiska institutionen, Umeå universitet.

Henriksson, W., Stage, C. & Lexelius, A. (1986). Samma uppgift men olika innehåll - En studie av NOG-provet. *PM från H-gruppen, 10*. Pedagogiska institutionen, Umeå universitet.

Stage, C. (1987). Analys av NOG-uppgifter med avseende på könsskillnader i resultat. *PM från H-gruppen, 39*. Pedagogiska institutionen, Umeå universitet.

Stage, C. (1988). Gender Differences in Test Results. *Scandinavian Journal of Educational Research, Vol 32, 3*, 101-111.

Wester-Wedman, A. (1992). Ett försök med öppna frågor i DTK-provet. En jämförelse mellan öppna frågor och flervalsfrågor avseende könsskillnaden i prestation på DTK-provet. *PM från H-gruppen, 55*. Pedagogiska institutionen, Umeå universitet.

Stage, C. (1992). Vad betyder form och innehåll för könsskillnader i provresultat? I *Betyg och högskoleprov för män och kvinnor*. UHÄ-rapport 1992:3.

Judgements of items with regard to gender differences.

Wedman, I. & Stage, C. (1983). The Significance of Contents for Sex Differences in Test Results. *Scandinavian Journal of Educational Research, Vol 27, 1.* 49-71.

Stage, C. (1985). En grupp prövandes bedömningar av vilka uppgifter som ger upphov till könsskillnader i resultat. I Stage, C. *Gruppskillnader i provresultat*. Akademisk avhandling nr 17. Pedagogiska institutionen, Umeå universitet. 119-133.

Emanuelsson, I. (1985). Könsskillnader och rättviseaspekter. *Forskning om utbildning 4*, 48-53. (Recension av Stages avhandling).

Stage, C. (1987). Skattning av könsskillnader i resultat på AO-uppgifter. *PM från H-gruppen, 11*. Pedagogiska institutionen, Umeå universitet.

Gender differences in different sub-groups of testtakers.

Stage, C. & Wedman, I. (1984). Lika möjligheter till utbildning. I *Utbildningsstatistisk årsbok 1983/84*. Stockholm: SCB. 33-46.

Stage, C. (1985). Ålders- och utbildningsskillnader i relation till könsskillnader i provresultat. I Stage, C. *Gruppskillnader i provresultat*. Akademisk avhandling nr 17. Pedagogiska institutionen, Umeå universitet. 135-154.

Stage, C. (1985). Resultatskillnader mellan män och kvinnor på samma prestationsnivå. I Stage, C. *Gruppskillnader i provresultat*. Akademisk avhandling nr 17. Pedagogiska institutionen, Umeå universitet. 155-173.

Stage, C. (1990). Könsskillnader i resultat på högskoleprovet våren 1990. *PM från H-gruppen, 42*. Pedagogiska institutionen, Umeå universitet.

Bränberg, K., Henriksson, W., Nyquist, H & Wedman, I. (1990) The Influence of Sex, Education and Age on Test Scores on the Swedish Scholastic Aptitude Test. *Scandinavian Journal of Educational Research, Vol 34, No 3.* 189-203.

Stage, C. (1991). Högskoleprovet våren 1991. Provdeltagargruppens sammansättning och resultat. *PM från H-gruppen, 48*. Pedagogiska institutionen, Umeå universitet.

Stage, C. (1992a). Högskoleprovet hösten 1991. Provdeltagargruppens sammansättning och resultat. *PM från H-gruppen, 59.* Pedagogiska institutionen, Umeå universitet.

Stage, C. (1992b). Gruppskillnader och resultat. I *Betyg och högskoleprov för män och kvinnor.* UHÄ-rapport 1992:3. 10-13.

Stage, C. (1992c). Högskoleprovet våren 1992. Provdeltagargruppens sammansättning och resultat. *PM från H-gruppen, 63.* Pedagogiska institutionen, Umeå universitet.

Stage, C. (1992d). How Important are Age and Education for Gender Differences in Test Results? *Scandinavian Journal of Educational Research, Vol 36, 3.* 223-235.

Stage, C. (1993). Högskoleprovet hösten 1992. Provdeltagargruppens sammansättning och resultat. *PM från H-gruppen, 72.* Avdelningen för pedagogiska mätningar, Umeå universitet.

Average School Marks and Test Results.

Stage, C. (1992e). Betyg och högskoleprov. *PM från H-gruppen, 53.* Pedagogiska institutionen, Umeå universitet.

Stage, C. (1992f). Skillnader mellan betyg och högskoleprovsresultat. *PM från H-gruppen, 62.* Avdelningen för pedagogiska mätningar, Umeå universitet.

Stage, C. (1992g). Average Marks and Test Results. Paper presented at the IAEA 18th Annual Meeting in Dublin, September, 1992.

Stage, C. (1993). Interaction between Gender, School Courses, Marks and Test Results. Unpublished manuscript, Division of Educational Measurement, University of Umeå.

EDUCATIONAL MEASUREMENT

Reports already published in the series

EM No 1. SELECTION TO HIGHER EDUCATION IN SWEDEN
Ingemar Wedman

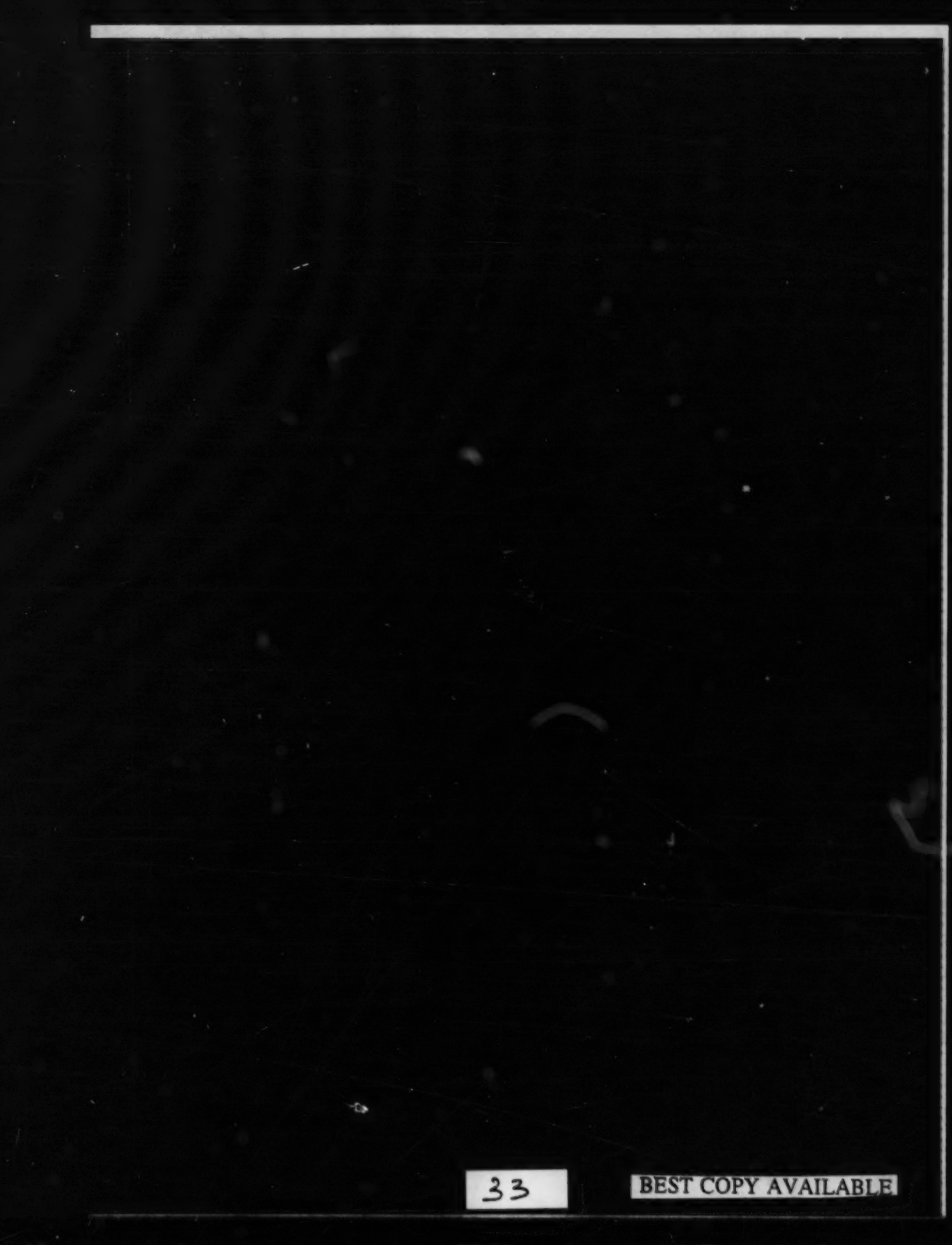
EM No 2. PREDICTION OF ACADEMIC SUCCESS IN A PERSPECTIVE OF CRITERION-RELATED AND CONSTRUCT VALIDITY
Widar Henriksson, Ingemar Wedman

EM No 3. ITEM BIAS WITH RESPECT TO GENDER INTERPRETED IN THE LIGHT OF PROBLEM-SOLVING STRATEGIES
Anita Wester

EM No 4. AVERAGE SCHOOL MARKS AND RESULTS ON THE SWESAT
Christina Stage

EM No 5. THE PROBLEM OF REPEATED TEST TAKING AND THE SweSAT
Widar Henriksson

EM No 6. COACHING FOR COMPLEX ITEM FORMATS IN THE SweSAT
Widar Henriksson



BEST COPY AVAILABLE